Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense

Year: 2023 | Citations: 421 | Authors: Kalpesh Krishna, Yixiao Song, Marzena Karpinska, J. Wieting, Mohit Iyyer

---

Abstract

The rise in malicious usage of large language models, such as fake content creation and academic plagiarism, has motivated the development of approaches that identify AI-generated text, including those based on watermarking or outlier detection. However, the robustness of these detection algorithms to paraphrases of AI-generated text remains unclear. To stress test these detectors, we build a 11B parameter paraphrase generation model (DIPPER) that can paraphrase paragraphs, condition on surrounding context, and control lexical diversity and content reordering. Using DIPPER to paraphrase text generated by three large language models (including GPT3.5-davinci-003) successfully evades several detectors, including watermarking, GPTZero, DetectGPT, and OpenAI's text classifier. For example, DIPPER drops detection accuracy of DetectGPT from 70.3% to 4.6% (at a constant false positive rate of 1%), without appreciably modifying the input semantics. To increase the robustness of AI-generated text detection to paraphrase attacks, we introduce a simple defense that relies on retrieving semantically-similar generations and must be maintained by a language model API provider. Given a candidate text, our algorithm searches a database of sequences previously generated by the API, looking for sequences that match the candidate text within a certain threshold. We empirically verify our defense using a database of 15M generations from a fine-tuned T5-XXL model and find that it can detect 80% to 97% of paraphrased generations across different settings while only classifying 1% of human-written sequences as AI-generated. We open-source our models, code and data.